# Data Standards for Data Analysis and Statistical Tools

- S.K. McWeeney & V. Rajaraman
- OHSU Cancer Institute
- CaBIG Face-to-Face Meeting
- August 24-25th, 2004

# Standards for Data Analysis and Statistical Tools

- Issues specific to this SIG:
  - This group is a consumer of data types addressed by other SIGS (microarray repositories, genome annotation etc)
  - Statistical analysis is downstream of the raw data that has been the focus of the data standard groups
  - However, there is a critical need for data standards for analysis and tools in order to facilitate data sharing, comprehension of results and allow full reproducibility.

- The few slides before our discussion will highlight similar issues from an MGED working group for statistical pre-processing and transformation that may serve as a starting point for us, as well as a brief look at some of the data standards that may be of use to this SIG.

# AN EXAMPLE FROM MGED

- **The MGED Data Transformation and Normalization Working Group**
  - This working group is focused on the development of recommendations regarding microarray data transformations and normalization methods.
- **Three Goals**
  - **Standardization**
  - **Metrics of Data Quality**
  - **Education**
- **To achieve these goals, key is community participation!**

# GOAL I: Standardization

- What does Standardization Imply?
  - It doesn't imply same transformations and normalization protocols must be used by everyone
  - Emphasis is on standards to record how data has been transformed and analyzed
    - This allows third party groups to assess whether methods were appropriate in context in which they were used
    - Also allows others to repeat exactly how analysis was done to determine if they reach the same conclusion

# How is this standardization achieved?

- Need to be able to describe transformation that may be carried out on data.
- Issue of how data is captured  (i.e., within MAGE-ML or extensions such as using MATH-ML for encoding equations or algorithms?)
- Need for ontologies, controlled vocabularies etc.

# For example: A Processing CV: What is needed

- A CV for what the main "processing operations" are (e.g., filtering, normalization, etc).
- Within each processing operation, a CV for the currently available methods. These CVs are aimed at capturing the main ideas of the methods, not the specific implementations.
- Specific implementation are illustrated by reference to a published or available algorithm or, when this is "ad hoc", a step by step description of the protocol.
- This processing CV would be used to record (1) what operations were done and in what order (2) the methods used and (3) for each method, the implementation used.

# GOAL 2: Metrics of Data Quality

- Metrics of Data Quality
  - Quality metrics were critical for sequence data. With the advent of good base-by-base quality metrics, assemblers were able to take data quality into account, and thus make much better assemblies. In the absence of such quality metrics, it is unlikely that accurate assembly of data genomic-scale sequencing would be feasible. **It is therefore not unreasonable to expect that accurate understanding and modeling of biological systems using genomic-scale expression data will at the very least benefit from data quality metrics.**
  - It is likely that there is  no single metric to describe quality of a given measurement and instead there may be many such metrics. Therefore, the working group is soliciting information from the microarray community about what metrics may be useful to determine if a spot or even the entire microarray is of good quality or not.

# Examples of Metrics in microarray studies

- Attributes of a spot that should be considered when determining quality:
  - How close is it to saturation?
  - How far above background is its signal?
  - How consistent is the measured ratio for each pixel in the spot?
  - How large is the spot?
- Determination of how reproducible a measurement is.
  - Multiple spots containing the same DNA sequence on the same microarray.
  - Multiple spots containing sequences that assay the same gene on the same microarray.
  - Replication of the hybridization or experiment, which may be at different levels of replication (e.g., hybridization, labeling, isolation)
- Metrics of array quality
  - Is there evidence of spatial bias?
  - What percentage of spots on the array are considered of good quality?
  - What is the overall signal and background like?

# GOAL 3: Education

- **Education/Dissemination**
  - Regardless of how simple or sophisticated a set of procedures for recording data transformations, and metrics of data quality that the working group devises, they will not be useful unless they are accepted and used by the microarray community.
  - This involves educating the community about what data transformation and normalization methods exist, as well as getting input from the community with regard to standards and metrics.
  - Ways to achieve this goal include tutorials, hosting materials on the MGED website, white papers etc.

# Other Efforts: Standardized Data/Analysis Transfer (Note: Not a comprehensive list by any means…)

- **Predictive Model Markup Language (PMML)**
  - http://www.dmg.org
  - Describes statistical and data mining models
- **Systems Biology Markup Language (SBML)**
  - http://sbml.org
  - A "tool-neutral" exchange format that is applicable to metabolic networks, cell-signaling pathways, genomic regulatory networks, and many other areas in systems biology.
- **Cell Markup Language (CellML)**
  - http://www.cellml.org
  - Designed to store and exchange computer-based mathematical models. Leverages MathML and RDF
- **Clinical Data Interchange Standards Consortium(CDISC) Analysis Data Modeling and Submissions Data Standard Team**
  - http://www.cdisc.org/standards/index.html
  - Goal of group is development of industry standards to support the electronic acquisition, exchange, submission and archiving of clinical trials data and metadata (including analysis) for medical and biopharmaceutical product development.

# Points for Discussion

- What are our starting and stopping points (e.g., with the MGED group, pre-processing refers to all transformations and statistical protocols after data is quantified but before it is analyzed)?
- What type of meta-data needs to be captured (what is the minimal amount of information for statistical analysis)?
- In order to capture this meta-data, will we need to first work with the Vocabularies and Common Data Elements workspace to establish new controlled vocabularies and/or ontologies?
- Which of the existing standards are available that can be utilized by the SIG (rather than reinventing the wheel)?
- Why do data standard efforts often fail? What can we learn from this?